Semi Adversarial Networks for Face De-identification

Arun Ross

Professor

Michigan State University

rossarun@cse.msu.edu

http://www.cse.msu.edu/~rossarun

The iPRoBe Lab

http://iprobe.cse.msu.edu **https://twitter.com/iPRoBeLab**

- Integrated Pattern Recognition and Biometrics Lab
- Currently: 7 PhD Students + 1 Post-Doc +2 UG Students
- Graduated: 24 MS Thesis Students + 7 PhD Students

Research Theme

- **Adversarial Biometric Recognition**
  - Spoofing Biometric Traits
  - Degraded Biometric Data
  - Heterogeneous Biometric Data
- **Forensics and Privacy**
  - What Else Does Your Biometric Data Reveal?
  - Privacy Preserving Biometrics
- **Biometric Fusion**
  - Multiple Biometrics
  - Biometrics + Demographics + Spoof Detector + Quality
  - Primary Biometrics + Soft Biometrics

National Science Foundation
WHERE DISCOVERIES BEGIN

IARPA
BE THE FUTURE

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

FEDERAL BUREAU OF INVESTIGATION
FBI BIOMETRIC Center of Excellence
Leading the World of Biometrics

ONR
OFFICE OF NAVAL RESEARCH

NIJ
NATIONAL INSTITUTE of JUSTICE

WV HIGHER EDUCATION POLICY COMMISSION

CITeR
CENTER FOR IDENTIFICATION TECHNOLOGY RESEARCH

BIMA
BIOMETRICS IDENTITY MANAGEMENT AGENCY

ARL

DRAPER LABORATORY

U.S. DEPARTMENT OF HOMELAND SECURITY

noblis
For the best of reasons

NIH National Institutes of Health
Turning Discovery Into Health

Related Papers

- V. Mirjalili, S. Raschka, A. Ross, "**Gender Privacy: An Ensemble of Semi Adversarial Networks for Confounding Arbitrary Gender Classifiers**," BTAS 2018
- V. Mirjalili, S. Raschka, A. Namboodiri, A. Ross, "**Semi-Adversarial Networks: Convolutional Autoencoders for Imparting Privacy to Face Images**," ICB 2018
- V. Mirjalili and A. Ross, "**Soft Biometric Privacy: Retaining Biometric Utility of Face Images while Perturbing Gender**," IJCB 2017
- A. Othman and A. Ross, "**Privacy of Facial Soft Biometrics: Suppressing Gender But Retaining Identity**," ECCVW 2014

CURRENT WORK FUNDED BY NATIONAL SCIENCE FOUNDATION

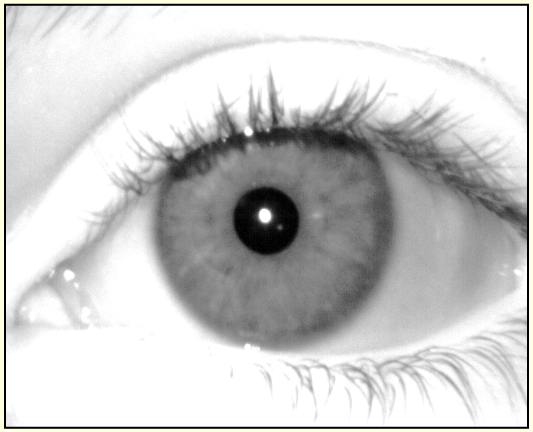
Privacy of Biometric Data

- Age, Gender, Ethnicity, can be **automatically derived** from the face image
- That is, a **trained classifier or a regressor** may be used to automatically deduce certain soft biometric attributes

- Gender: Male
- Age: 25
- Health: Very good
- Eye Sight: Wears glasses
- Ethnicity: Asian Indian

Dantcheva, Elias, Ross, "What Else Does Your Biometric Data Reveal? A Survey on Soft Biometrics," TIFS 2016

Biometrics + Forensics



- Subject is a Male (90% Confidence), White (85% Confidence)
- Image taken using an Aoptix camera
- Iris stroma is plain textured
- Highly constricted pupil suggests strong ambient illumination

Bridges the gap between human and machine description of data
OR
Compromises privacy?

Surveillance Applications

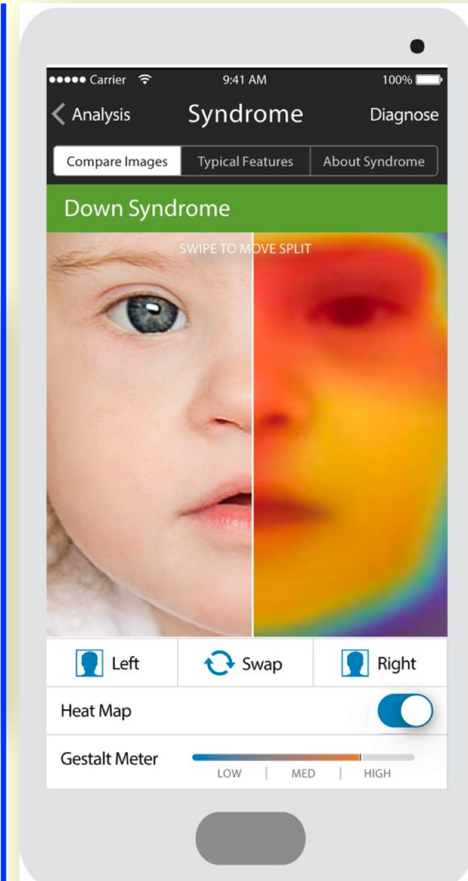
Face2Gene

MEGAN MOLTENI SCIENCE 01.09.17 01:00 PM

THANKS TO AI, COMPUTERS CAN NOW SEE YOUR HEALTH PROBLEMS

"In hindsight it was all clear to me," says Gripp, who is chief of the Division of Medical Genetics at A.I. duPont Hospital for Children in Delaware, and had been seeing the patient for years. "But it hadn't been clear to anyone before." What had taken Patient Number Two's doctors 16 years to find took Face2Gene just a few minutes.

Face2Gene is a suite of phenotyping applications that facilitate comprehensive and precise genetic evaluations.

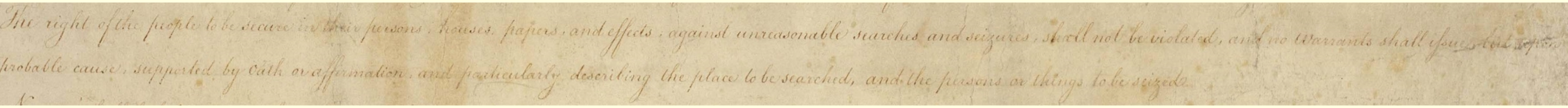


Identifying People on the Web

- **Faces of Facebook: Privacy in the Age of Augmented Reality (Alessandro Acquisti)**
- Convergence of three technologies:
  - face recognition, cloud computing, online social networks
- Started from an anonymous face in the street
- Ended up with very sensitive information about that person → **data accretion**
- Combined face recognition with the algorithms they developed in 2009 to predict SSNs from public data

Importance of Privacy

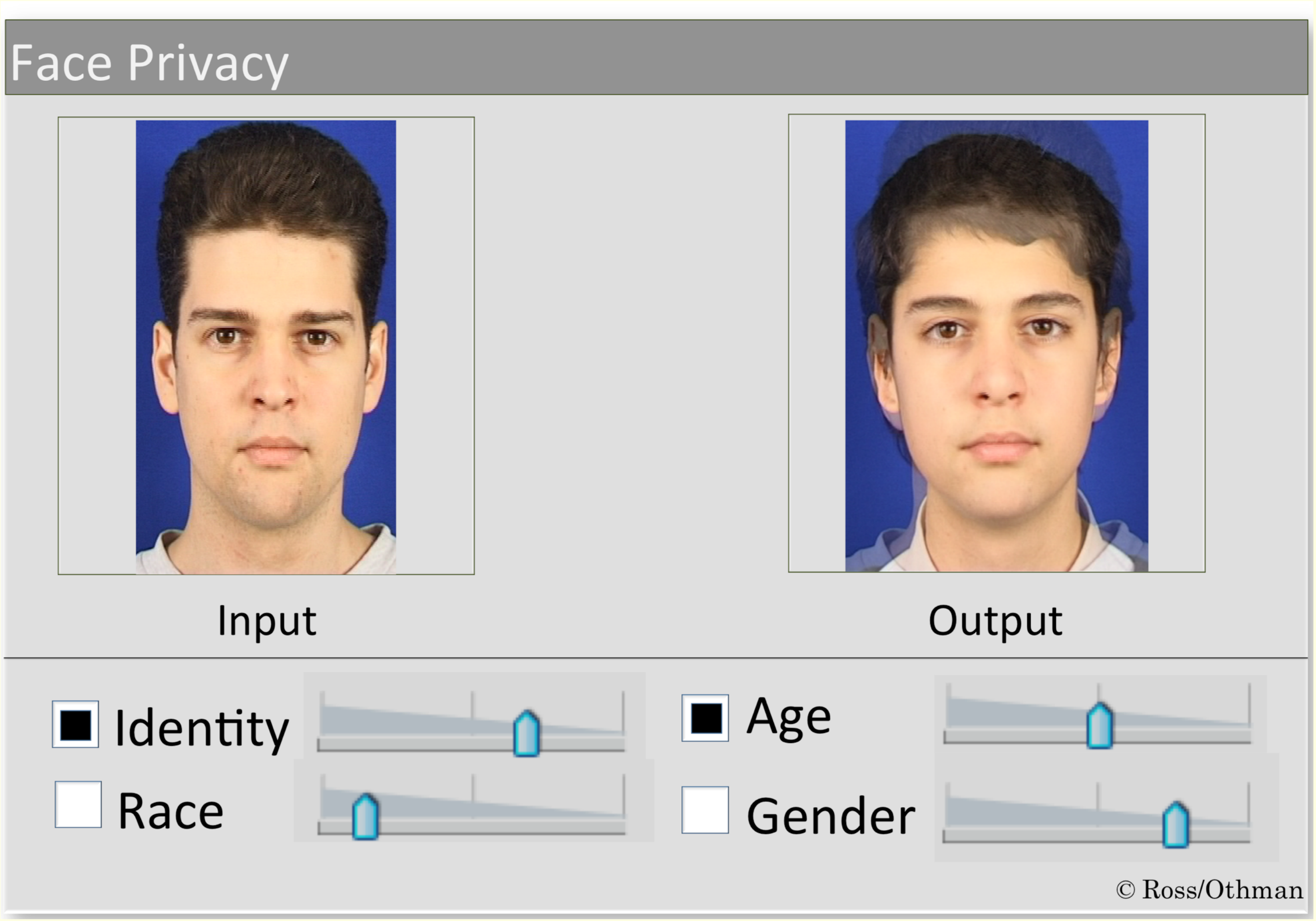
- “Privacy is the right to be let alone” [Samuel Warren and Louis Brandeis (1890)]
- “Privacy is the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others” [Alan Westin (1970)]
- “Privacy is the right of people to conceal information about themselves that others might use to their disadvantage” [Richard Posner (1983)]


The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no Warrants shall issue,
probable cause, supported by Oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized.

PRIVACY IS DIFFERENT FROM SECURITY

“Differential” Privacy

Differential Privacy



Othman and Ross, "Privacy of Facial Soft Biometrics," ECCVW 2014

Differential Privacy

- We investigate the possibility of preserving the contextual integrity of face images stored in a central biometric database
- We consider the problem of suppressing a soft biometric attribute of a face
- This modification should not drastically impact the accuracy of the automated face matcher

Soft Biometric Privacy

- Gender attribute of an input face image is progressively suppressed
- With respect to a face matcher the recognition capability is preserved

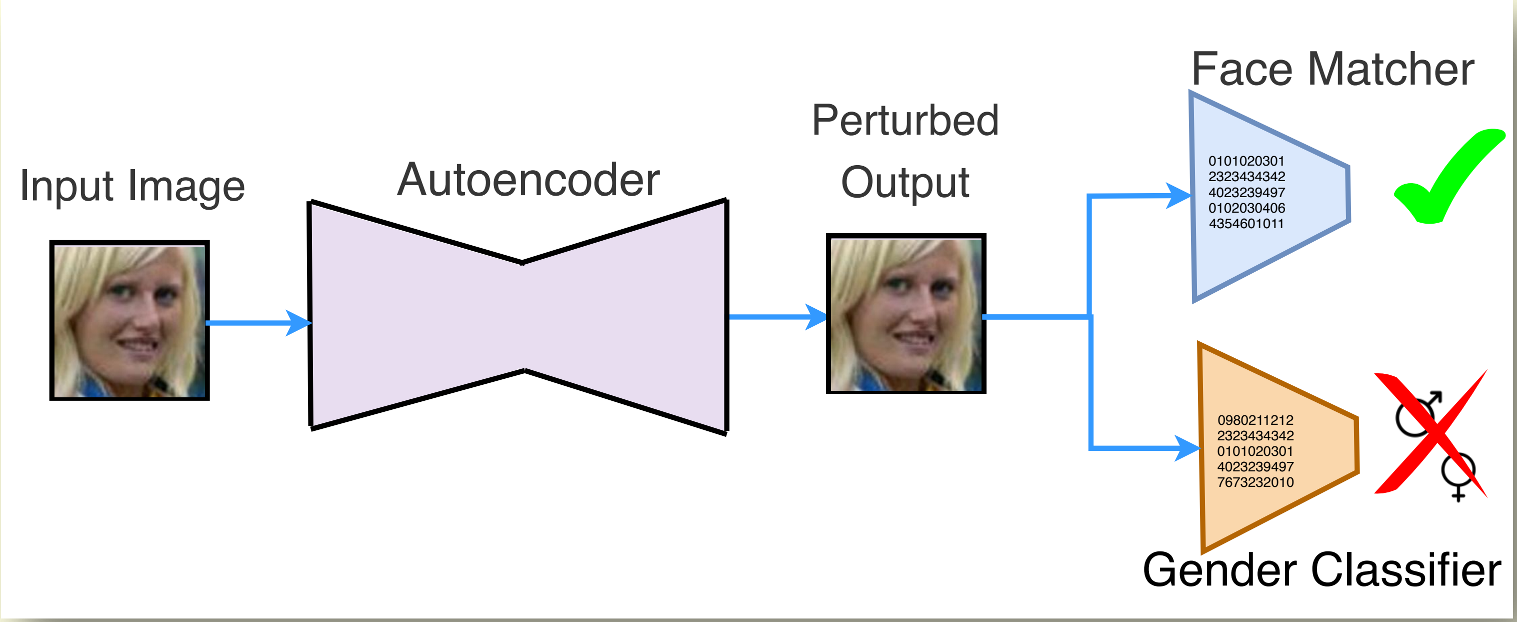
Input image Transformed images

Name	Alice	Alice	Alice	Alice
Gender	Female (confident)	Female (less confident)	Male (less confident)	Male (confident)

Othman and Ross, "Privacy of Facial Soft Biometrics: Suppressing Gender But Retaining Identity", ECCV Workshop, 2014

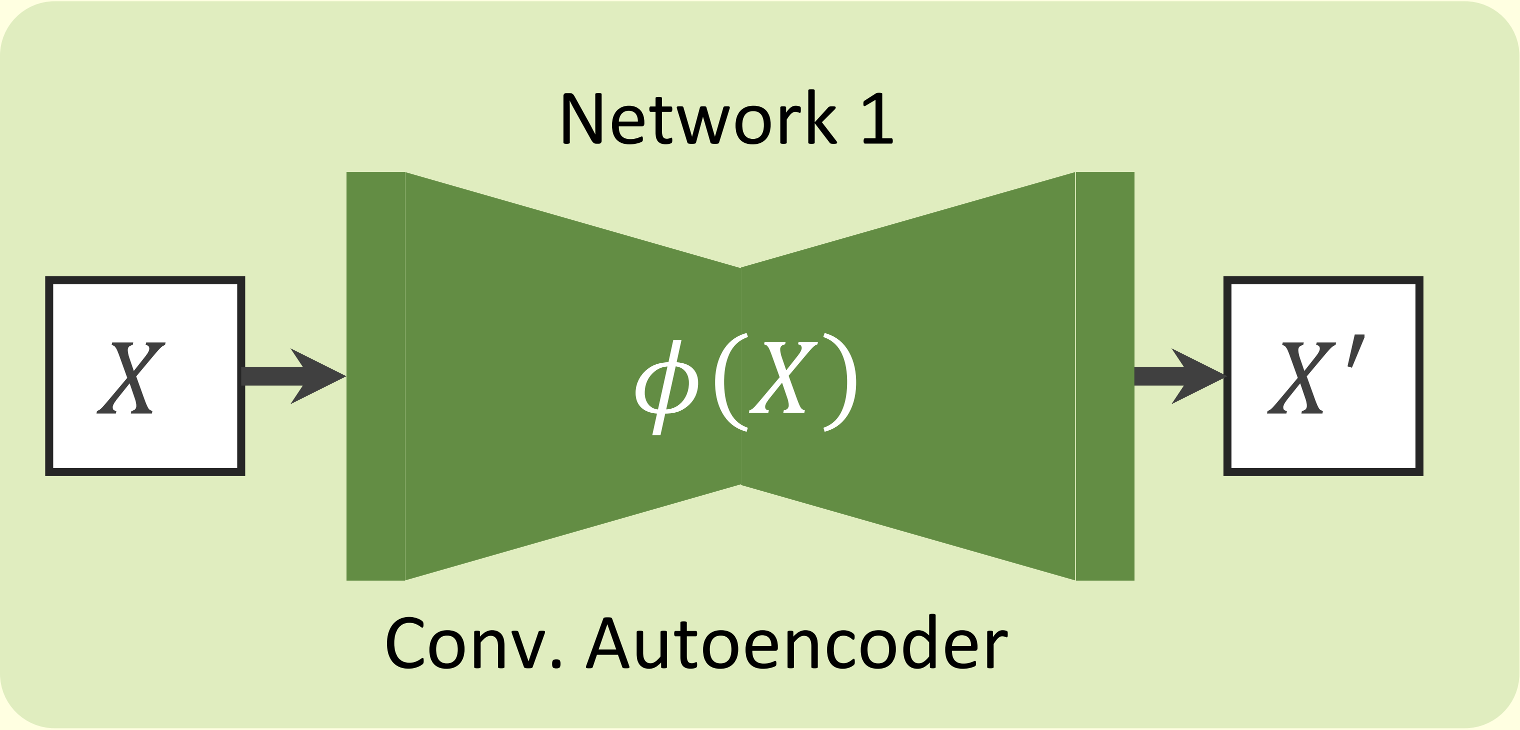
Semi-Adversarial Networks (SAN)

- Design a transformation model to:
  - Confound gender attribute ➔ gender classifiers will not work
  - Retain recognition capability ➔ face matchers will still work



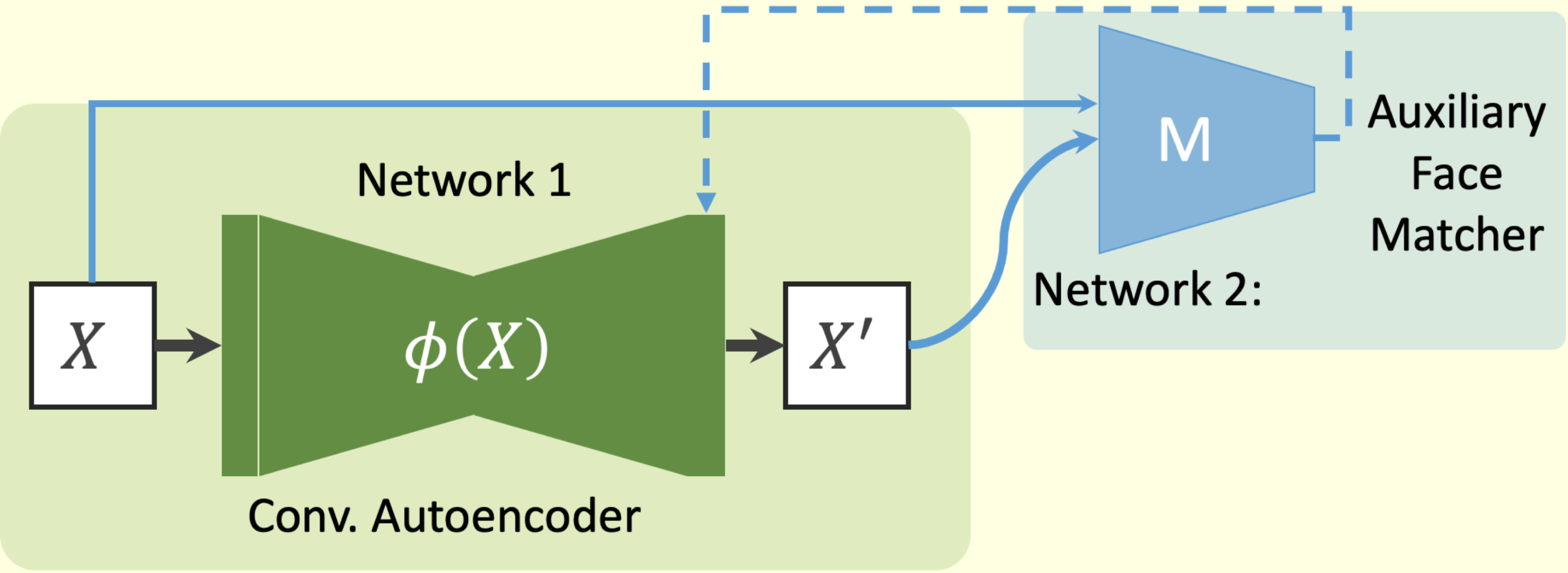
Mirjalili et al., Semi-Adversarial Networks: Convolutional Autoencoders for Imparting Privacy to Face Images, ICB 2018

General Architecture of SAN Model



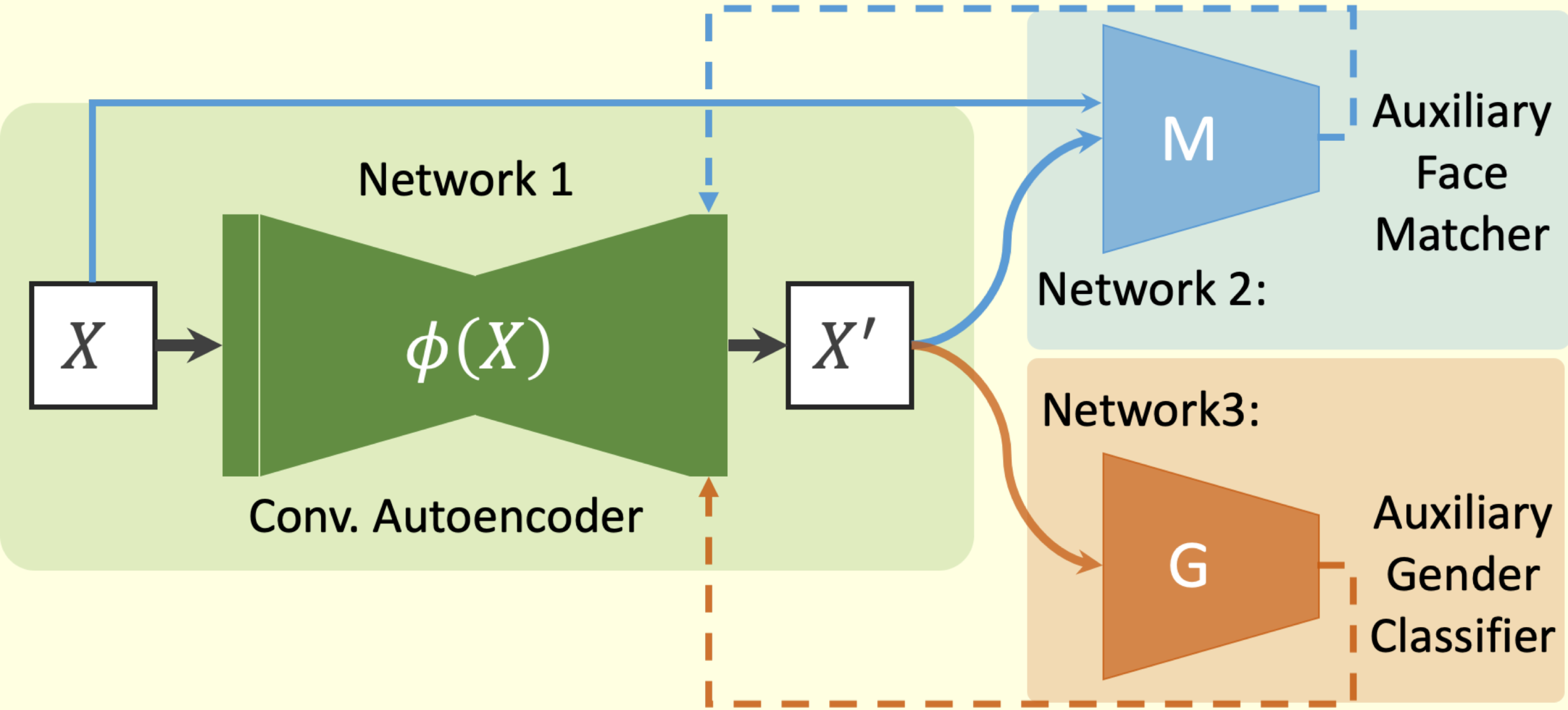
Mirjalili et al., Semi-Adversarial Networks: Convolutional Autoencoders for Imparting Privacy to Face Images, ICB 2018

General Architecture of SAN Model



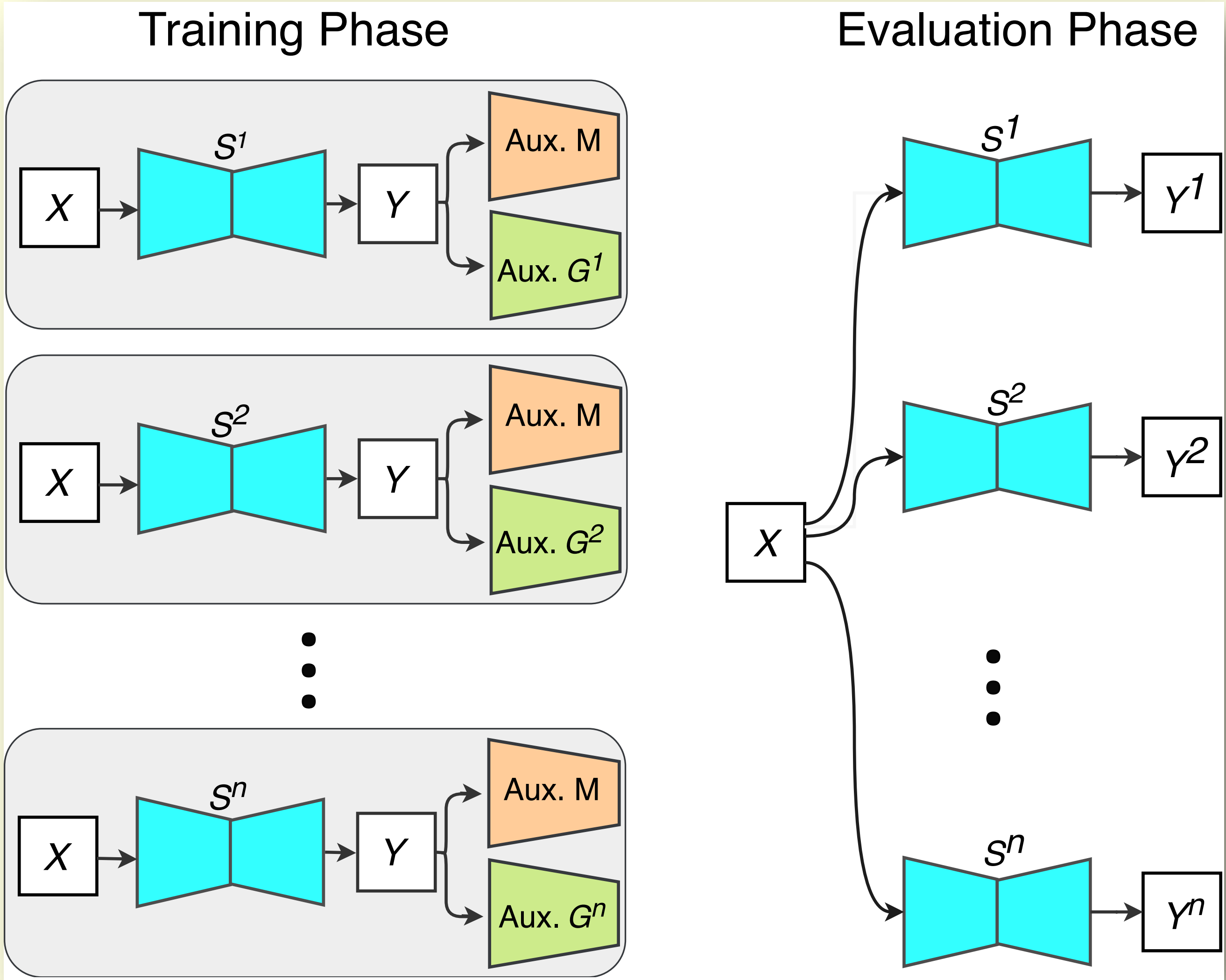
Mirjalili et al., Semi-Adversarial Networks: Convolutional Autoencoders for Imparting Privacy to Face Images, ICB 2018

General Architecture of SAN Model



Mirjalili et al., Semi-Adversarial Networks: Convolutional Autoencoders for Imparting Privacy to Face Images, ICB 2018

Ensemble of SANs



V. Mirjalili, S. Raschka, A. Ross, "Gender Privacy: An Ensemble of Semi Adversarial Networks for Confounding Arbitrary Gender Classifiers," BTAS 2018

Cost Functions for Semi-Adversarial Learning

1. **Pixel-wise similarity term**

$$J_D(X, X'_{SM}) = \sum_{k=1}^{N} S\left(X^{(k)}, X'^{(k)}_{SM}\right)$$

   - Only used during the pre-training of Autoencoder

2. **Loss term related to gender attribute**
   - Correctly predict gender of X'_{SM}
   - Flip the gender prediction on X'_{OP}

$$J_G(X, X'_{SM}, X'_{OP}, y; f_G) = S\big(y, f_G(X'_{SM})\big) + S\big(1 - y, f_G(X'_{OP})\big)$$

3. **Loss term related to face identity matching**

$$J_M\big(X, X'_{SM}; R_{vgg}\big) = \left|\left|R_{vgg}(X'_{SM}) - R_{vgg}(X)\right|\right|_2^2$$

Training Protocol

- **Auxiliary subnetworks**
  - Auxiliary gender predictor is trained on CelebA dataset, and its parameters are frozen during training of Conv. Autoencoder
  - Publicly available parameters for VGG are used for the auxiliary face matcher
- **Training the Autoencoder**

Step1: pre-training the Conv. Autoencoder with two loss terms: pixel-wise similarity + gender term

Step2: replace the pixel-wise similarity term with the matching term based on VGG subnetwork (trained for 20 epochs)

Examples of Inputs and Outputs

Male: 99%

Female: 98%

Male: 97%

Male: 100%

Female: 69%

Male: 99%

Male: 71%

Female: 58%

Mirjalili et al., Semi-Adversarial Networks: Convolutional Autoencoders for Imparting Privacy to Face Images, ICB 2018

Examples of Inputs and Outputs




Male: 98%

Male: 99%

Female: 100%

Female: 99%




Female: 79%

Female: 53%

Male: 63%

Male: 67%

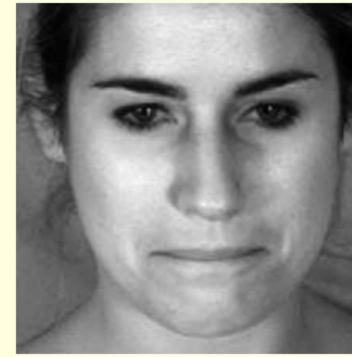
Mirjalili et al., Semi-Adversarial Networks: Convolutional Autoencoders for Imparting Privacy to Face Images, ICB 2018

Examples of Inputs and Outputs

Male:
100%

Male:
85%

Female:
100%



Female:
99%

Female:
95%

Female:
51%

Male:
75%

Male:
78%

Mirjalili et al., Semi-Adversarial Networks: Convolutional Autoencoders for Imparting Privacy to Face Images, ICB 2018

Examples of Inputs and Outputs

Male:
99%

Male:
88%

Male:
99%

Male:
94%

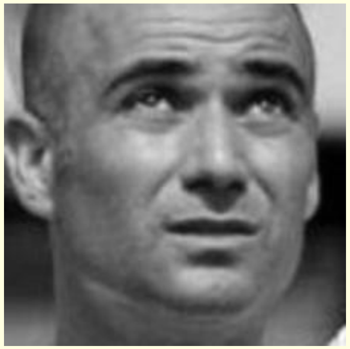
Male:
52%

Female:
91%

Female:
56%

Female:
93%

Examples of Inputs and Outputs




Male: 98%

Female: 72%

Female: 94%

Female: 99%




Male: 85%

Female: 80%

Male: 95%

Male: 52%

Mirjalili et al., Semi-Adversarial Networks: Convolutional Autoencoders for Imparting Privacy to Face Images, ICB 2018

Datasets Statistics

Dataset	# Samples	# Subjects	# Male Images	# Female Images
CelebA-train	157,350	--	65,160	92,190
CelebA-test	39,411	--	16,318	23,093
MUCT	3,754	276	1,844	1,910
LFW	12,988	5,658	10,083	2,905
AR-face	3,286	136	1,821	1,465

- CelebA dataset was split into train and test
- CelebA-train was used for training the autoencoder as well as the auxiliary gender predictor

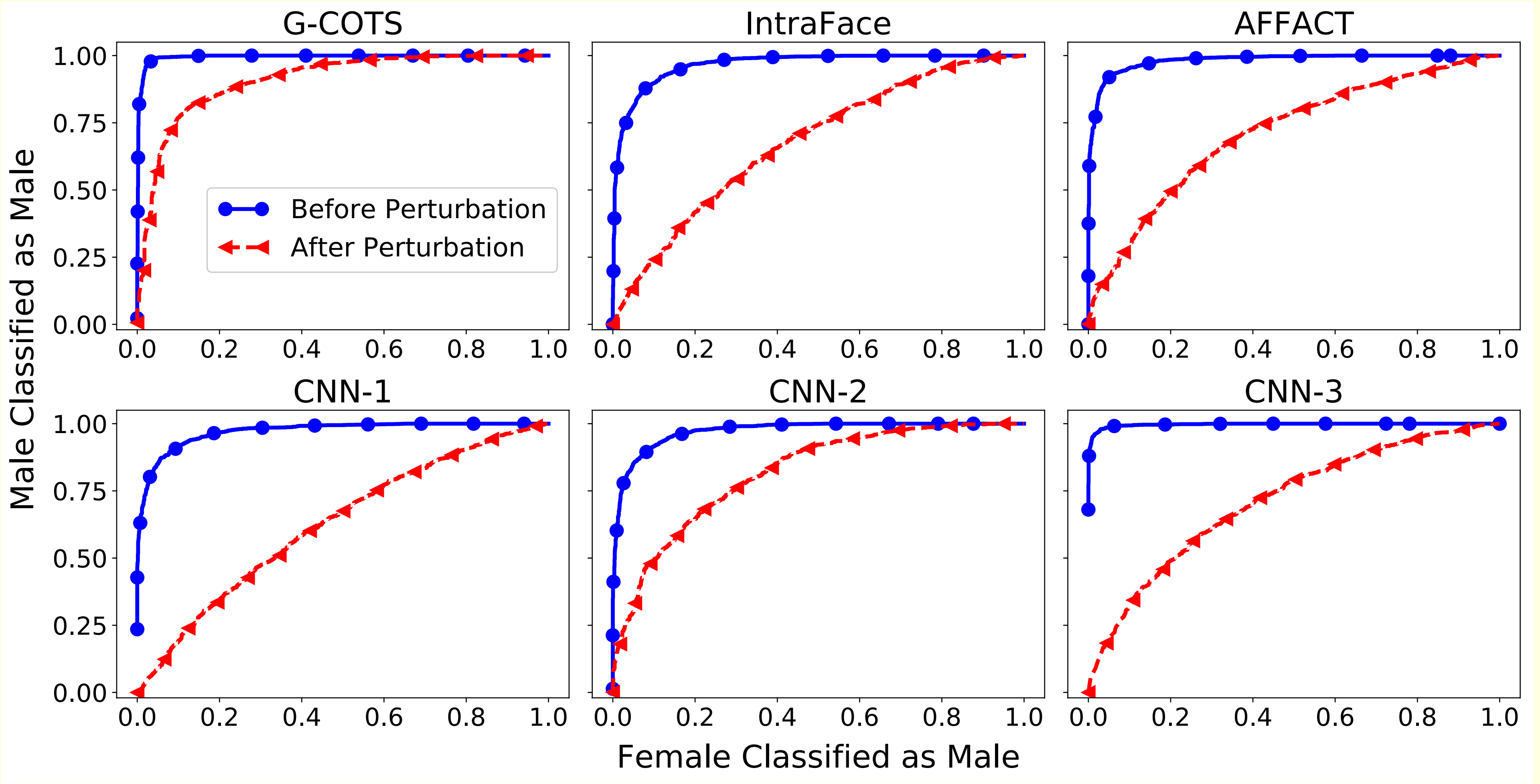
Experimental Design

- Six unseen gender Classifiers
  - G-COTS [Commercial]
  - IntraFace [De la Torre et al., 2015]
  - AFFACT [Günther et al., 2017]
  - 3 CNN models [in-house]

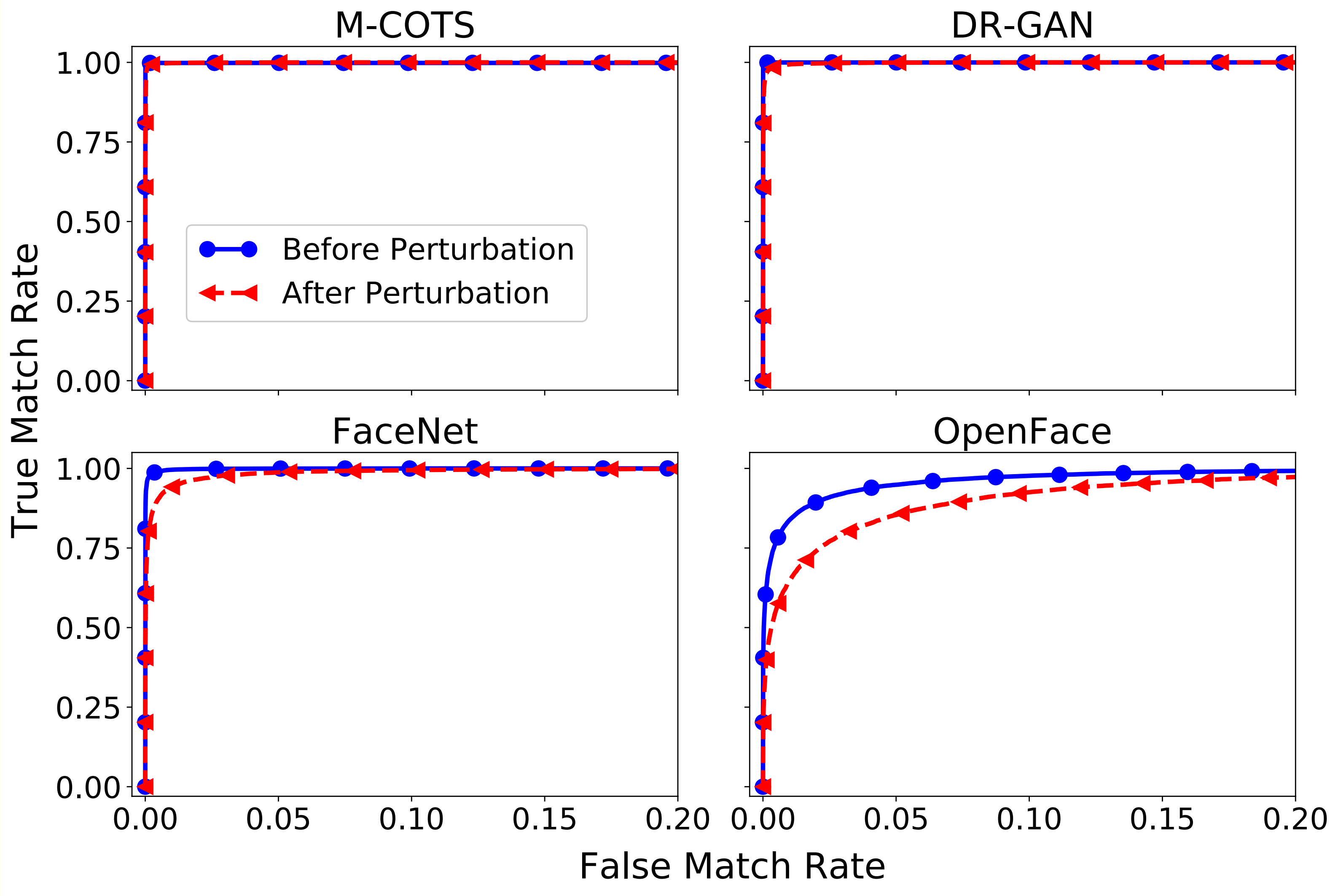
- Four unseen face Matchers
  - M-COTS [Commercial]
  - DR-GAN [Tran et al., 2017]
  - FaceNet [Schroff et al., 2015]
  - OpenFace [Amos et al., 2016]

Unseen:
the classifier or face matcher is <u>not used during training</u> of the SAN models

Performance Assessment on MUCT dataset: Confound gender classifiers



Performance Assessment on MUCT dataset: Retain Matching Capability



Ross/2018

Summary

- **Semi-Adversarial Network**
  - Perturbing one classifier while retaining the performance of other
- **Results confirm that**
  - Automatic gender prediction is confounded ➔ providing gender privacy to face images
  - Matching utility is still retained
- **Future work**
  - Extending to multiple attributes: gender, age, ethnicity
  - Differential privacy: some attributes preserved, others confounded
  - Visual realism of images

Mirjalili et al., Semi-Adversarial Networks: Convolutional Autoencoders for Imparting Privacy to Face Images, ICB 2018

Privacy Enhancing Technology

- Preserving the privacy of a user's stored biometric data
  - Regulate cross-linking across applications
  - Regulate gleaning additional information from biometric data (e.g., medical condition)

Need to
- Define Privacy and Privacy Metrics
- Guarantee Privacy
- Develop Differential Privacy Schemes

Semi Adversarial Networks for Face De-identification

Arun Ross

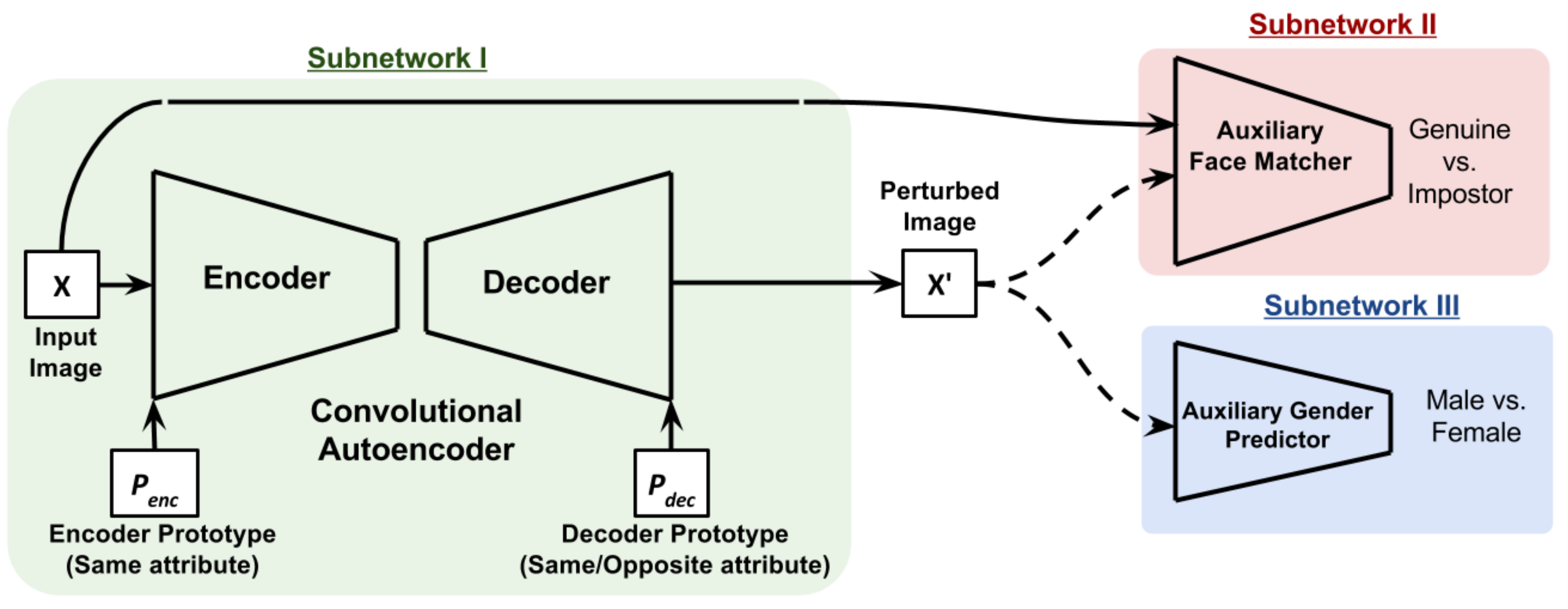
Professor

Michigan State University

rossarun@cse.msu.edu

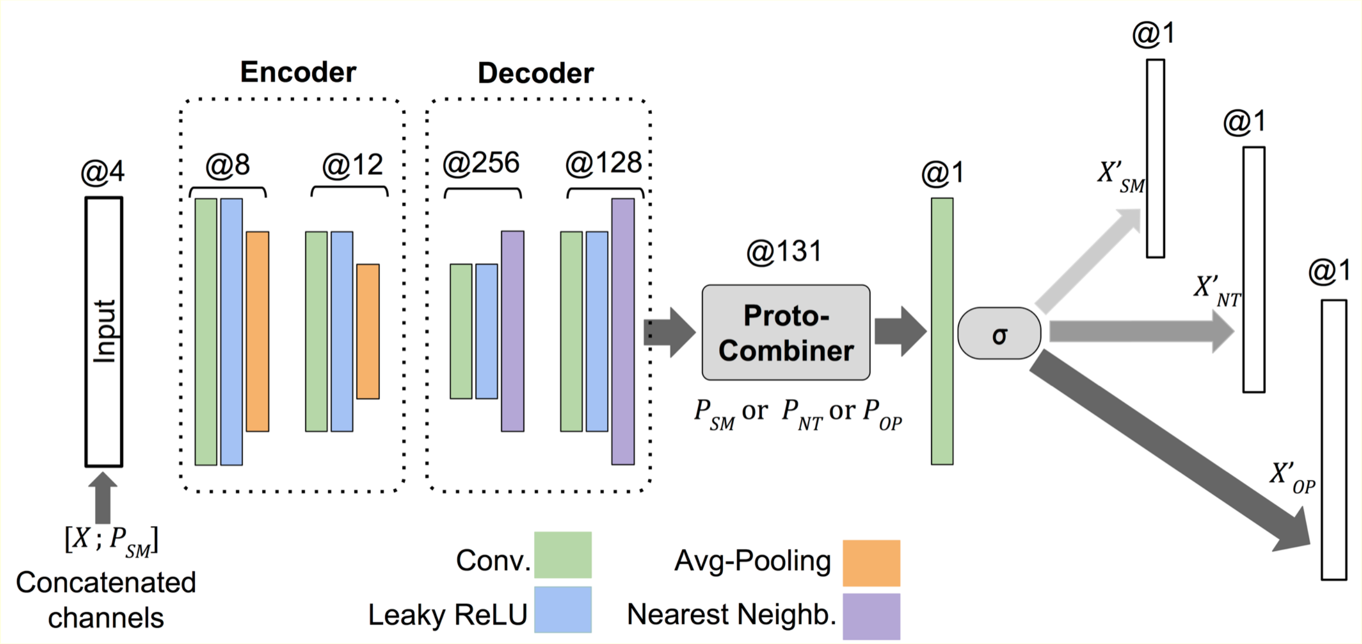
http://www.cse.msu.edu/~rossarun

General Architecture of SAN Model



Mirjalili et al., Semi-Adversarial Networks: Convolutional Autoencoders for Imparting Privacy to Face Images, ICB 2018

Subnetwork I: Convolutional Autoencoder

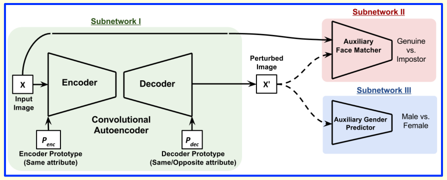


- With the use of different prototypes, three different outputs are generated:

  X'_{SM}: gender is not confounded

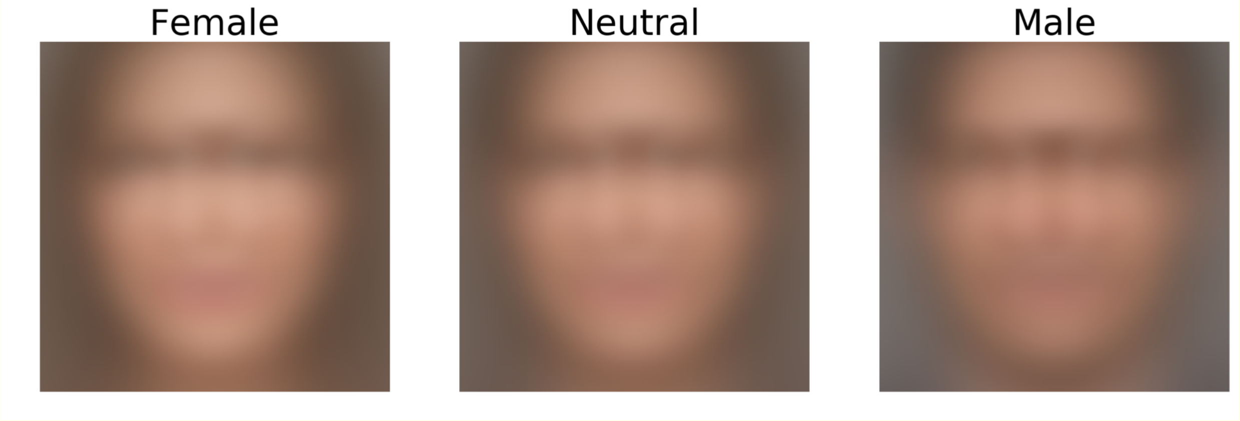
  X'_{OP}: gender is confounded

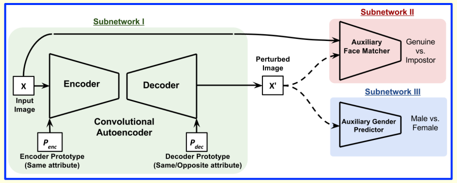
  X'_{NT}: middle-ground



Face Attribute Prototypes

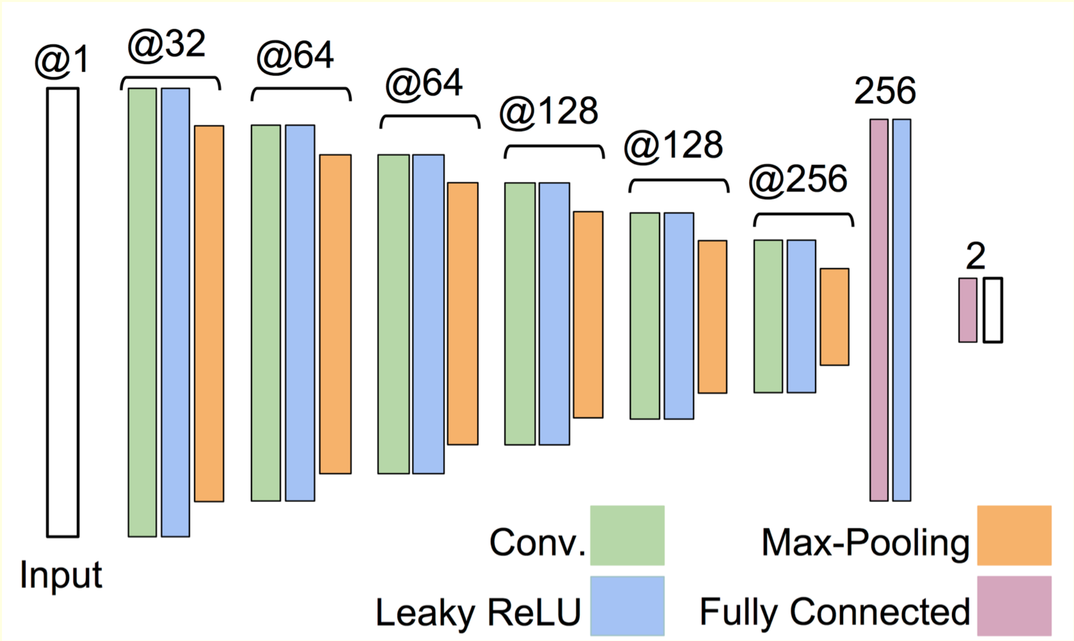
- Gender prototypes are computed as the mean image from both male and female faces:
  - P_{Male}: average of male images
  - P_{Female}: average of female images
  - $P_{Neutral}$: average of all images



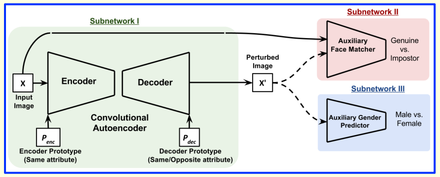


Ross/2018

Subnetwork II: Auxiliary Gender Predictor



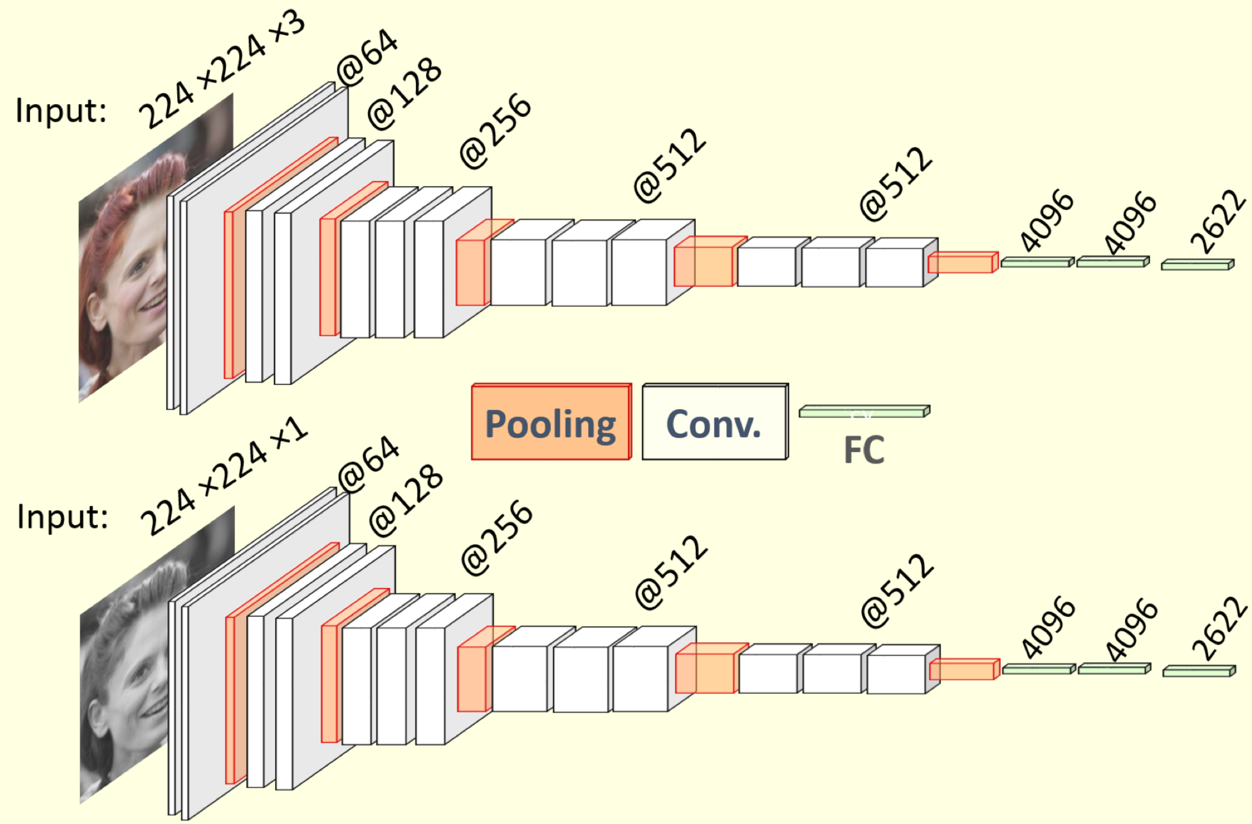
- Six convolution layers followed by a fully-connected layer and *softmax*



Subnetwork III: Auxiliary Face Matcher

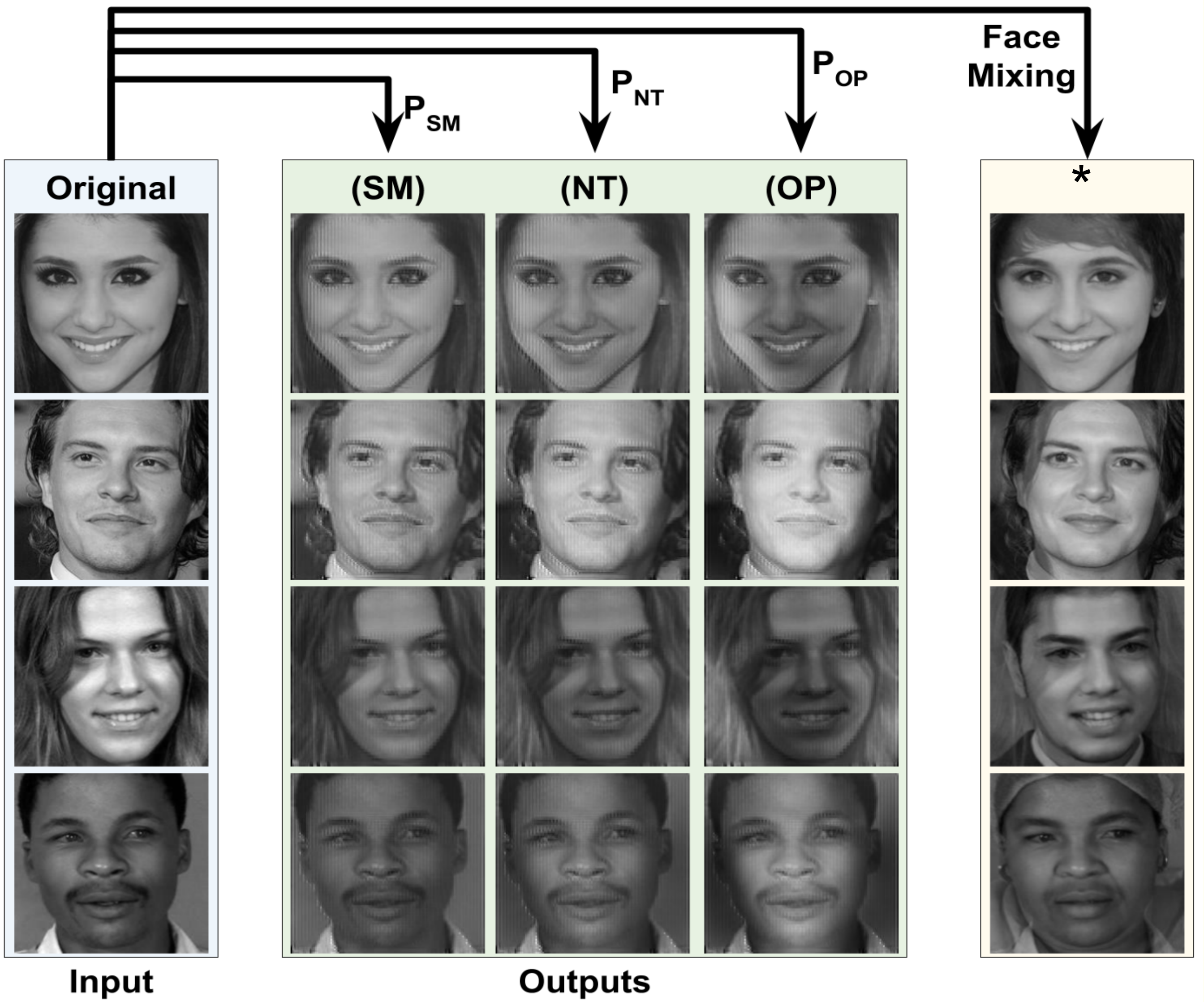
Original VGG network for RGB input:

Modified VGG network for gray-scale input:



- 16 weight layer
- Output: Face representation of size 2622

Examples of Inputs and Outputs



Gender Prediction Error Rates

Performance of G-COTS

Dataset	Original (before)	Perturbed (after OP)
CelebA-test	19.7%	39.3%
MUCT	8.0%	39.2%
LFW	33.4%	72.5%
AR-face	16.9%	53.8%

Performance of IntraFace

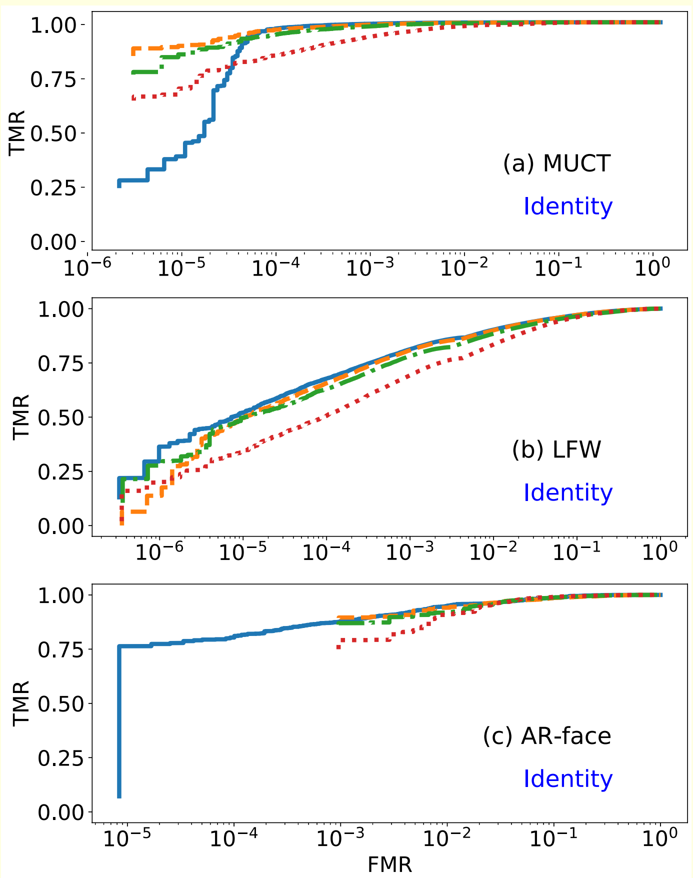
Dataset	Original (before)	Perturbed (after OP)
CelebA-test	19.7%	39.3%
MUCT	8.0%	39.2%
LFW	33.4%	72.5%
AR-face	16.9%	53.8%

Increase in **gender prediction** error rates confirms that *automatic* gender prediction is confounded

Mirjalili et al., Semi-Adversarial Networks: Convolutional Autoencoders for Imparting Privacy to Face Images, ICB 2018

Performance in Retaining Matching

ROC curves of match-scores obtained from M-COTS



TMR values at FMR=0.01

TMR : True Match Rate
FMR : False Match Rate

Dataset	Original (before)	Perturbed (after)		
		Same	Neutral	Opposite
MUCT	99.88	99.79	99.57	98.44
LFW	90.29	90.02	88.47	83.45
AR-face	94.97	94.11	91.95	90.81

The result verifies that the **matching accuracy** is **NOT** unduly affected by the perturbations

Mirjalili et al., Semi-Adversarial Networks: Convolutional Autoencoders for Imparting Privacy to Face Images, ICB 2018